Simulation of the entire range of daily precipitation using a hybrid probability distribution

Chao Li,[1] Vijay P. Singh,[1,2] and Ashok K. Mishra[1]

Received 26 September 2011; revised 8 February 2012; accepted 15 February 2012; published 21 March 2012.

[1] Underestimation of extreme values is a widely acknowledged issue in daily precipitation simulation. Nonparametric precipitation generators have inherent limitations in representing extremes. Parametric generators can realistically model the full spectrum of precipitation amount through compound distributions. Nevertheless, fitting these distributions suffers from numerical instability, supervised learning, and computational demand. This study presents an easy-to-implement hybrid probability distribution to model the full spectrum of precipitation amount. The basic idea for the hybrid distribution lies in synthesizing low to moderate precipitation by an exponential distribution and extreme precipitation by a generalized Pareto distribution. By forcing the two distributions to be continuous at the junction point, the threshold of the generalized Pareto distribution can be implicitly learned in an unsupervised manner. Monte Carlo simulation shows that the hybrid distribution is capable of modeling heavy tailed data. Performance of the distribution is further evaluated using 49 daily precipitation records across Texas. Results show that the model is able to capture both the bulk and the tail of daily precipitation amount. The maximum goodness-of-fit and penalized maximum likelihood methods are found to be reliable complements to the maximum likelihood method, in that generally they can provide adequate goodness-of-fit. The proposed distribution can be incorporated into precipitation generators and downscaling models in order to realistically simulate the entire range of precipitation without losing extreme values.

Citation: Li, C., V. P. Singh, and A. K. Mishra (2012), Simulation of the entire range of daily precipitation using a hybrid probability distribution, *Water Resour. Res.*, *48*, W03521, doi:10.1029/2011WR011446.

1. Introduction

[2] Precipitation simulation is one of the key features of hydrological models, agricultural models, and climate impact studies [*Kleiber et al.*, 2011]. Sufficiently long series of precipitation records are needed for catchment water management, drought characterization and prediction, and crop growth simulation. However, historical records of precipitation with desired spatial and temporal resolution are almost always insufficient. Moreover, it is difficult to quantify the uncertainty of model results from only a single sequence of realizations. On the other hand, there is considerable discussion these days that climate change is contributing to the increase in frequencies and magnitudes of precipitation extremes, leading to floods or droughts, and hence evaluating changes in precipitation extremes is receiving significant attention [*Solomon et al.*, 2007; *Lenderink and Meijgaard*, 2008; *Hardwick Jones et al.*, 2010]. Therefore, realistically modeling the full spectrum of precipitation is desired.

[3] Precipitation simulation dates back to the 1950s. Over the past decades, many simulation techniques have been developed [e.g., *Gabriel and Neumann*, 1962; *Katz*, 1974, 1977; *Todorovic and Woolhiser*, 1975; *Richardson*, 1981; *Stern and Coe*, 1984; *Lall and Sharma*, 1996; *Wilks*, 1998; *Rajagopalan and Lall*, 1999; *Parlange and Katz*, 2000; *Yan et al.*, 2002; *Harrold et al.*, 2003a, 2003b; *Chandler*, 2005; *Mehrotra and Sharma*, 2007a, 2007b; *Furrer and Katz*, 2007; *Zheng and Katz*, 2008a, 2008b; *Brissette et al.*, 2007]. Typically, daily precipitation is represented as a mixture of two distributions in a parametric, nonparametric, or semiparametric framework. One is discrete binary distribution modeling the wet or dry state of a given day, and the other is continuous distribution modeling nonzero precipitation amounts on wet days. A most recent review on precipitation simulation can be found by *Sharma and Mehrotra* [2010]. Overall, there are two acknowledged challenges in daily precipitation simulation. One is referred to as overdispersion [*Katz and Zheng*, 1998]. The other one is the loss of extreme precipitation events. The first problem concerns both the occurrence and amount processes of precipitation, whereas the second one mainly concerns the amount process. This paper focuses on the second problem.

[4] Since daily precipitation amount always shows a skewed distribution with a bias toward low values, it is usually modeled by distribution families which have right-skewed property [*Hundecha et al.*, 2009]. Different

[1]Department of Biological & Agricultural Engineering, Texas A&M University, College Station, Texas, USA.
[2]Department of Civil & Environmental Engineering, Texas A&M University, College Station, Texas, USA.

0043-1397/12/2011WR011446

distributions, such as Kappa [*Mielke*, 1973], exponential [*Todorovic and Woolhiser*, 1975; *Roldan and Woolhiser*, 1982], gamma [*Ison et al.*, 1971; *Katz*, 1977; *Schoof et al.*, 2010], mixed exponential [*Roldan and Woolhiser*, 1982; *Wilks*, 1998, 1999], and truncated and power transformed normal distributions [*Bárdossy and Plate*, 1992; *Hutchinson*, 1995] have been used to model daily precipitation amount. The aforementioned families perform reasonably well in terms of reproducing averaging characteristics of precipitation. Nevertheless, none of them necessarily performs well in terms of simulating extremes [*Wilks*, 1999; *Furrer and Katz*, 2008]. Besides parametric approaches, nonparametric approaches also have been used for daily precipitation simulation. Synthetic precipitation is sequentially sampled from historical observations with replacement. Several limitations, especially with respect to extremes, inherent to the sampling scheme [*Furrer* and *Katz*, 2008], have been recognized, and corrected via nonparametric kernel density estimator (KDE) [*Lall and Sharma*, 1996; *Rajagopalan and Lall*, 1999]. Nevertheless, the likelihood for extremes to be generated is low [*Markovich*, 2007], leading to underestimated extreme rainfall. Reproducing the entire range of precipitation in synthetic series has been identified as a critical research need in both simulation and downscaling, and has inspired a recent flurry of research, like *Vrac and Naveau* [2007], *Furrer and Katz* [2008], and *Hundecha et al.* [2009], in which compound distributions are used for modeling precipitation amount. Problems involved in fitting these distributions include numerical instability, data sensitivity, supervised learning, and computational demand.

[5] The objective of this study therefore is to develop an efficient, reliable and relatively easy-to-implement method for simulating both the low to moderate and extreme rainfall. To that end, the specific objectives are to: (1) examine if the existing distributions are reliable to model precipitation amount in different climate divisions, (2) present a hybrid distribution to model the full spectrum of daily precipitation, (3) validate the hybrid distribution model, and (4) describe approaches for estimating parameters of the hybrid distribution. It is noted that the proposed distribution is not intended, however, to replace existing distributions, like those developed by *Vrac and Naveau* [2007] and *Furrer and Katz* [2008], but rather to complement them and to provide a more efficient, reliable, and less complicated approach to model the heavy tail distribution of precipitation amount without losing any goodness-of-fit.

[6] The paper is organized as follows. Formulating the objectives of the study in section 1, a short discussion of data to be used is given in section 2. Fundamental to the simulation of daily rainfall is the choice of a probability distribution which is described in section 3. A hybrid probability distribution is presented in section 4. Its evaluation is presented in section 5. Based on problems raised from real cases, three estimation approaches are presented in section 6. The paper is concluded in section 7.

2. Data Sets

[7] Daily precipitation records from 49 weather stations across Texas, given in Table 1, were used. These stations are spread across 10 climate divisions which are divided by National Weather Service. The United States Historical Climatology Network (USHC) provides high quality precipitation records [*Mishra and Singh*, 2010]. This study concerns only precipitation amount. Therefore, all nonzero records were valued pieces of information. Missing values have little influence on fitted distributions as long as sufficient data were available. To gather as much useful information as

Table 1. ID, Label, Location, Annual Mean Precipitation (P) and Temperature (T) of Weather Stations in Texas Used in This Study[a]

Station ID	Station Label	Longitude (deg)	Latitude (deg)	P (mm)	T (°F)	Station ID	Station Label	Longitude (deg)	Latitude (deg)	P (mm)	T (°F)
410120	ID1	32.73	−99.30	700.28	63.15	413734	ID26	33.17	−96.09	1094.23	65.51
410144	ID2	27.73	−98.07	707.64	72.26	413873	ID27	29.47	−96.94	1035.03	67.84
410174	ID3	30.38	−103.66	412.75	61.97	413992	ID28	33.16	−99.75	649.73	62.87
410493	ID4	31.74	−99.98	597.41	64.77	415018	ID29	31.07	−98.18	787.40	67.25
410498	ID5	30.98	−103.74	333.50	63.90	415196	ID30	30.06	−94.79	1434.85	68.44
410639	ID6	28.46	−97.71	822.45	69.58	415272	ID31	30.74	−98.65	695.71	65.48
410832	ID7	30.11	−98.43	879.35	65.39	415429	ID32	29.68	−97.66	919.48	67.61
410902	ID8	29.79	−98.73	922.72	65.77	415618	ID33	32.54	−94.35	1244.35	64.57
411000	ID9	35.53	−102.26	428.91	57.43	415707	ID34	31.13	−102.22	359.16	66.23
411048	ID10	30.16	−96.39	1117.10	66.41	415869	ID35	31.70	−96.51	1013.21	66.07
411138	ID11	31.74	−98.95	713.23	64.58	415875	ID36	35.70	−100.64	572.01	57.54
411528	ID12	28.34	−99.63	462.28	70.32	416135	ID37	34.22	−102.73	446.53	56.71
411772	ID13	33.62	−95.07	1170.94	63.09	416276	ID38	29.72	−98.12	858.01	68.21
412015	ID14	27.77	−97.51	790.96	71.57	416794	ID39	33.67	−95.56	1172.46	64.66
412019	ID15	32.11	−96.47	986.03	66.17	416892	ID40	31.42	−103.50	273.56	64.77
412121	ID16	33.65	−101.25	567.18	60.48	417079	ID41	34.19	−101.70	510.29	59.36
412266	ID17	29.06	−96.23	1136.65	68.38	417336	ID42	34.28	−99.76	614.93	61.89
412598	ID18	32.06	−98.30	874.27	63.98	417622	ID43	26.38	−98.81	565.15	74.05
412679	ID19	28.76	−100.48	531.37	71.61	417945	ID44	29.53	−98.47	795.78	68.53
412797	ID20	31.81	−106.38	223.77	64.21	418201	ID45	32.71	−102.65	447.80	62.07
412906	ID21	28.02	−99.35	533.15	72.79	418433	ID46	32.71	−100.91	562.36	62.66
413063	ID22	27.14	−98.12	641.09	72.08	418692	ID47	36.34	−102.08	441.19	55.42
413183	ID23	29.68	−97.11	967.49	70.03	418910	ID48	31.08	−97.32	908.30	64.52
413280	ID24	30.91	−102.92	340.44	65.16	419532	ID49	32.75	−97.77	858.27	62.64
413420	ID25	33.65	−97.06	993.39	62.96						

[a]Annual mean precipitation and temperature are computed using data over 1960 to 2009.

possible, all nonzero precipitation from the period of 1940 to 2009 was used without taking care of missing values. Omitting the influence of missing values is harmless considering that there were at least 1600 nonzero records at each station.

3. Evaluation of Commonly Used Distributions for Daily Precipitation

[8] Fundamental to the simulation of daily precipitation is the use of an appropriate probability distribution for precipitation amount. To that end, the question arises: What are the typical distributional characteristics of daily precipitation? Then, one should search for a distribution that captures these characteristics. Since there are several forms of possible distributions, the next question to be addressed is one of evaluating these distributions and selecting the appropriate one. The selection of a distribution involves enumerating distribution properties and estimation of distribution parameters. These issues are discussed in what follows.

3.1. Rainfall Characteristics

[9] First, typical characteristics of a nonzero daily precipitation distribution were explored, using, as an example, the central station ID24, which is located in the western part of Texas. Among all the 49 stations, ID24 had the maximum entropy and was hence considered as "central" [*Krstanovic and Singh*, 1992]. Without considering missing values, there were 3049 wet days from 1 May 1940 to 31 December 2009. A histogram of nonzero values, together with summary statistics, is shown in Figure 1(a), which exhibits a representative shape of the distribution of daily precipitation amount. Two typical properties seen from the histogram include: (1) right skewed, indicated by median being lower than mean and most of the data being clustered around the left end of the distribution; and (2) heavy tailed, represented by sparse observations toward the tail end of the distribution and slower than exponential decay to zero, which can be efficiently illustrated from Figure 1(b). Distributions which can simulate these two properties should be used. Extensively employed distributions can generally be divided into single and compound types.

3.2. One-Component Distributions

[10] Commonly used one-component distributions include Kappa [*Mielke*, 1973], exponential [*Todorovic and Woolhiser*, 1975; *Woolhiser and Roldan*, 1982], and gamma distributions [*Ison et al.*, 1971; *Katz*, 1977; *Schoof et al.*, 2010]. Let X denote the nonzero daily precipitation amount and subscript capital letters, say K (for Kappa distribution), to distinguish different distributions. The probability density functions (PDF) for these distributions are now presented.

[11] Kappa distribution:

$$f_K(x;\mathbf{P}_K) = \frac{a\theta}{b}\left(\frac{x}{b}\right)^{\theta-1}\left[a+\left(\frac{x}{b}\right)^{a\theta}\right]^{-\frac{a+1}{a}}, \quad x>0, \quad a,b,\theta>0, \tag{1}$$

where $\mathbf{P}_K = [a, b, \theta]$, and a, b, and θ are the shape and scale parameters, respectively.

[12] Exponential distribution:

$$f_E(x;P_E) = \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right), \quad x\geq 0, \quad \mu>0, \tag{2}$$

where $P_E = \mu$ and μ is the scale or intensity parameter.

[13] Gamma distribution:

$$f_G(x;\mathbf{P}_G) = \frac{1}{b\Gamma(a)}\left(\frac{x}{b}\right)^{a-1}\exp\left(-\frac{x}{b}\right), \quad x\geq 0, \quad a,b>0, \tag{3}$$

where $\mathbf{P}_G = [a, b]$, and a and b are the shape and scale parameters, respectively.

[14] Other distributions where the focus is on modeling only the upper tail include generalized stretched exponential distribution and generalized Pareto (GP) distribution [*Coles*, 2001; *Katz et al.*, 2002; *Koutsoyiannis*, 2004a, 2004b; *Wilson and Toumi*, 2005; *Naveau et al.*, 2005]. Since the objective in this study was to simulate not only the "tail" but also the "bulk," our interest is only in distributions which are widely used in stochastic weather generators or those that can model both the aforementioned precipitation distribution properties.

3.3. Two-Component Distributions

[15] Commonly used compound distributions include mixed exponential distribution [*Roldan and Woolhiser*, 1982; *Wilks*, 1998, 1999], dynamic mixture of gamma and

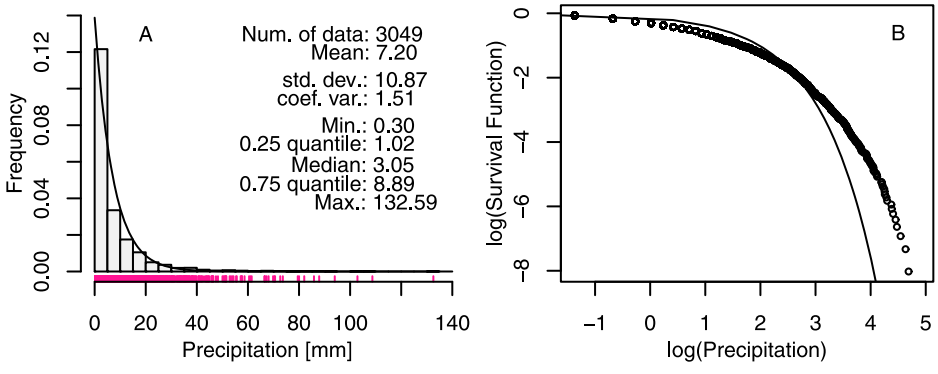

Figure 1. Histogram of precipitation amount of station ID24 with the fitted exponential PDF represented by (a) solid line and the empirical survival function with that of the fitted exponential distribution denoted by (b) solid line plotted on log-log scale.

generalized Pareto distribution [*Vrac and Naveau*, 2007; *Hundecha et al.*, 2009], and hybrid gamma and generalized Pareto distribution [*Furrer and Katz*, 2008]. These compound distributions are discussed below.

[16] Mixed exponential distribution (ME):

$$f_{\mathrm{ME}}(x;\mathbf{P}_{\mathrm{ME}}) = p\frac{1}{\mu_1}\exp\left(-\frac{x}{\mu_1}\right) + (1-p)\frac{1}{\mu_2}\exp\left(-\frac{x}{\mu_2}\right), \quad x \geq 0, \quad \mu_1, \mu_2 > 0, \quad p \in [0,1], \tag{4}$$

where $\mathbf{P}_{\mathrm{ME}} = [p, \mu_1, \mu_2]$, p is the mixing factor, and μ_1, μ_2 are, respectively, the scale parameters of the two components.

[17] Dynamic mixture of gamma and GP distribution (DM):

$$f_{\mathrm{DM}}(x;\mathbf{P}_{\mathrm{DM}}) = \frac{[1-p(x;\mu,\tau)]f_G(x;a,b) + p(x,\mu,\tau)f_{\mathrm{GP}}(x;\kappa,\sigma,0)}{Z(a,b,\mu,\tau,\kappa,\sigma)}, \quad x > 0, \quad a,b,\mu,\tau,\kappa,\sigma > 0, \tag{5a}$$

where $\mathbf{P}_{\mathrm{DM}} = [a,b,\mu,\tau,\kappa,\sigma]$, and $p(x;\mu,\tau)$ is the mixing function expressed as

$$p(x,\mu,\tau) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\tau}\right) \tag{5b}$$

with location parameter μ and scale parameter τ. The mixing function monotonically increases from 0.5 to 1 as x increases from 0 to ∞ such that the bulk of the distribution is dominated by gamma and the tail is dominated by the GP distribution.

[18] The other ingredients in this distribution are the gamma density $f_G(x;a,b)$ parameterized by a and b, the GP density $f_{\mathrm{GP}}(x;\kappa,\sigma,\theta)$ located at 0 ($\theta = 0$) with shape parameter κ and scale parameter σ, and the normalization constant $Z(a,b,\mu,\tau,\kappa,\sigma)$. After easy algebraic operations, the normalization constant is given as

$$Z(a,b,\mu,\tau,\kappa,\sigma) = 1 + \frac{1}{\pi}\int_0^\infty [f_{\mathrm{GP}}(x;\kappa,\sigma,0) - f_G(x;a,b)] \arctan\left(\frac{x-\mu}{\tau}\right)dx. \tag{5c}$$

The advantages of this DM distribution are: (1) it can model the full range of precipitation; and (2) it circumvents the selection of threshold. Threshold selection is a challenging task in practice since given a data set it is difficult to pinpoint the level on which the extreme value theory (EVT) is based and usually a subjective trial and error exercise is needed [*Frigessi et al.*, 2002; *Carreau and Bengio*, 2009].

[19] Hybrid gamma and GP distribution:

[20] Considering the discontinuity of DM distribution in the limiting case (τ is 0 or close to 0) and the difficulty of incorporating covariates, *Furrer and Katz* [2008] proposed a hybrid distribution where a GP distribution replaces the tail of a gamma distribution. For simplicity we use FK08 to denote this distribution. The PDF of FK08 distribution is

$$f_{\mathrm{FK08}}(x;\mathbf{P}_{\mathrm{FK08}}) = f_G(x;a,b)I(x \leq \theta) + [1 - F_G(\theta;a,b)]f_{\mathrm{GP}}(x;\kappa,\sigma,\theta)I(x > \theta), \tag{6a}$$

where $I(\cdot)$ is the indicator function and $1 - F_G(\theta;a,b)$ is the normalization factor. This factor ensures that the integral of the density over its support is unity. To force the hybrid density to be continuous at the threshold θ, it is necessary that $f_{\mathrm{FK08}}(\theta-) = f_{\mathrm{FK08}}(\theta+)$, which yields that the scale parameter σ of the GP distribution is exactly the reciprocal of the gamma hazard function, i.e.,

$$\sigma = \frac{1 - F_G(\theta;a,b)}{f_G(\theta;a,b)}. \tag{6b}$$

Therefore, this distribution can be fully represented by $\mathbf{P}_{\mathrm{FK08}} = [a,b,\kappa,\theta]$.

[21] In both the above compound distributions, the GP distribution acts as a tail component

$$f_{\mathrm{GP}}(x;\mathbf{P}_{\mathrm{GP}}) = \frac{1}{\sigma}\left(1 + \kappa\frac{x-\theta}{\sigma}\right)^{-\frac{1}{\kappa}-1}, \quad x > \theta, \quad 1 + \kappa\frac{x-\theta}{\sigma} > 0, \quad \sigma > 0, \tag{7}$$

where $\mathbf{P}_{\mathrm{GP}} = [\kappa,\sigma,\theta]$, and κ, σ, and θ are the shape, scale, and location parameters, respectively. The popularity of GP distribution can be explained by EVT [*Coles*, 2001; *Castillo et al.*, 2005], which states that precipitation exceedances over a threshold θ can be asymptotically approximated by a GP distribution, given that the threshold is sufficiently large. EVT has a deficiency of overlooking small values since the threshold should be sufficiently high and since only exceedances are involved in the analysis. Therefore, it is not suitable for modeling the full spectrum of precipitation. Then it would seem intuitive to model the bulk of precipitation by gamma or exponential distribution and take care of the tail by GP.

3.4. Estimation of Distribution Parameters

[22] The maximum likelihood (ML) method was used to estimate distribution parameters. To estimate parameters of the DM distribution, the normalization constant should be computed and the log likelihood function should be maximized. Since there is no closed antiderivative for the integral in the normalization constant, numerical integration should be employed. In this study, the MATLAB function *quadgk*, which is based on the adaptive Gauss-Kronrod quadrature algorithm, was used to complete the numerical integration. The log likelihood function was minimized with the use of *fminsearch*, which implements the Nelder-Mead simplex method. To obtain highly precise normalization constant, *Frigessi et al.* [2002] suggested to rewrite the integral from 0 to ∞ as a sum of integrals from 0 to 1, 1 to 2, and so on, compute each integral and truncate the summation when the last integral did not lead to any significant change in the sum. This approach is computationally expensive since numerical experiments show that generally a large number of steps are needed and since the normalization constant should

be computed at each step of the log likelihood maximization procedure. To speed up the procedure, an alternative is to compute the integration directly from 0 to ∞ or to a large value and then use *Frigessi*'s method. The fitting approach works satisfactorily when the integrand behaves well, as shown in the upper plot of Figure 2, but deteriorates, however, when the integrand is not smooth, as shown in the bottom plot. In this case, a normalization constant with low accuracy is obtained, which in turn introduces false local maxima of the log likelihood function and finally confounds the optimization algorithm.

[23] It is difficult to determine the FK08 distribution directly by the ML method. This distribution was therefore estimated following the procedure suggested by *Furrer and Katz* [2008]: (1) estimating the gamma parameters a and b by the ML method with all the data; (2) determining a reasonable threshold θ and then estimating the GP distribution κ and σ by the ML method from the data above θ; and (3) adjusting the estimated scale parameter σ obtained in step 2 from equation (6b) to achieve a continuous density. The estimation procedure underscores that a reasonable threshold should be predetermined.

3.5. Evaluation of Single Component Distributions Using QQ Plots

[24] The ML method was used to fit the aforementioned distributions to precipitation data from a sample station ID24. QQ plot was chosen as a goodness-of-fit criterion. QQ plots corresponding to Kappa, exponential, gamma distributions, given in Figure 3, show that despite an acceptable fit for low to moderate values, all of these distributions provided a rather poor fit for higher values. The upper tails of exponential and gamma distributions are not heavy enough, thus underestimating the likelihood of heavy precipitation, whereas Kappa distribution exhibits a much too heavy tail, leading to the overestimated extreme precipitation.

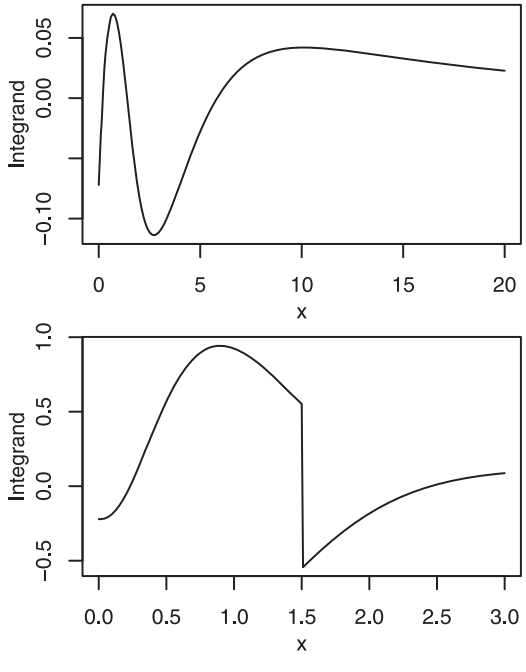

Figure 2. Two representative behaviors of the integrand function in the normalization constant of the DM distribution.

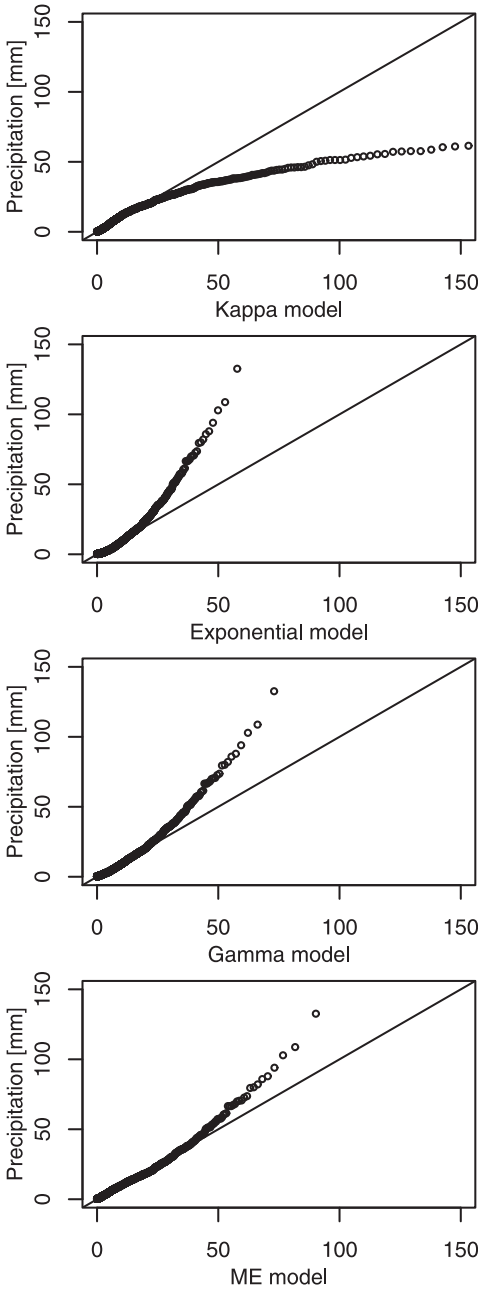

Figure 3. QQ plots of observed versus Kappa, exponential, gamma, and ME modeled precipitation quantiles of station ID24.

3.6. Evaluation of Compound Distributions Using QQ Plots

[25] As seen from Figure 3, the ME distribution, which is the most extensively used compound distribution for precipitation simulation, offered somewhat of an improvement in capturing the tail behavior but not enough. This observation empirically confirms that the ME distribution performs well only when precipitation extremes are not very high [*Wilks*, 1999; *Hundecha et al.*, 2009].

[26] Since the quantile function of the DM distribution cannot be analytically expressed and since we want to use the QQ plot to check the agreement between data and the fitted distribution, a parametric bootstrap method was used

to construct the QQ plot [*Efron and Tibhirani*, 1993; *Gomes and Oliveira*, 2001; *Castillo et al.*, 2005]. First, a large number of samples were drawn from the fitted distribution and the corresponding estimates for quantile $x_{i:n}$ were computed. These estimates were then used to obtain an empirical cumulative distribution function (CDF) for the estimated quantiles $\hat{x}_{i:n}$. The bootstrap based QQ plot can be constructed as a scatterplot of $E(\hat{x}_{i:n})$ versus $x_{i:n}$, and the corresponding confidence interval, say 95%, can be computed as $[\hat{x}_{i:n}(0.025), \hat{x}_{i:n}(0.975)]$.

[27] In the quest to achieve this goal, one problem still to be addressed is how to simulate random samples from the DM distribution. Due to its functional complexity, direct random number simulation method is no longer feasible. A simulation approach, introduced by *Frigessi et al.* [2002], was therefore employed. The step by step procedure and the pseudo code are given in the Appendix.

[28] Figure 4 presents the QQ plots and their 95% confidence intervals for three sample stations, ID11, ID24, and ID44, respectively. The figure reveals two main points: (1) similar to commonly used single component distributions, the DM distribution provides an adequate fit for the "bulk" of precipitation amount, and (2) its performance on capturing high values depends on the data. It models the full range of precipitation well at station ID11, but for other two stations it leads to an over heavy tail, which may be caused by the distributional property of data (as in station ID24) or by the low accuracy of the normalization constant (as in station ID44). Although the DM distribution can be a viable choice to model the full spectrum of precipitation, its application is constrained by several problems, such as functional complexity, numerical instability, and computational expense.

[29] QQ plots of the fitted FK08 distribution corresponding to different thresholds are presented in Figure 5, signifying that its performance is determined by the threshold, which should be neither too small nor too large. A too small threshold (for example, $\theta = 1.02$) means over emphasis on the GP distribution which will lead to an over heavy tail; whereas a too large threshold (say, $\theta = 18.19$) indicates less emphasis on the GP distribution which will result in an underrepresented tail. A suitable threshold (like, $\theta = 3.99$) does model both low to moderate and extreme values well. The threshold should be manually selected by trial and error, which is laborious and oftentimes subjective to the preference of a practitioner. Take $\theta = 3.99$ and $\theta = 18.19$ as examples, peaks over threshold (PoT) analysis indicates that both values are reasonable to model the exceedances by the GP distribution, as shown from the two plots in Figure 5. However, the fitted FK08 distributions significantly differ from each other in their performance with respect to full range of modeling.

4. Proposed Hybrid Distribution

[30] The above discussion shows that (1) a single distribution is inadequate to model the full range of daily precipitation and (2) fitting the available compound distributions suffer from functional complexity, numerical instability, supervised learning, and computational demand. For daily precipitation simulation, both the bulk and the tail should be taken into account. On the other hand, a computationally efficient model is attractive. To simulate the full range of precipitation, it is desirable to circumvent the threshold selection and reduce model complexity without loss of ability, if any. Taking into account these considerations, this study chose between the DM and the FK08 distributions to build a hybrid distribution by coupling an exponential distribution and a GP distribution. The presented hybrid distribution has its origin in the one introduced by *Carreau and Bengio* [2009], where Gaussian and GP distributions were stitched together.

[31] The PDF of the hybrid exponential and GP distribution (HEG) is given as

$$f_{\mathrm{HEG}}(x;\mathbf{P}_{\mathrm{HEG}}) = \frac{1}{Z}[f_E(x;\mu)I(x \le \theta) + f_{\mathrm{GP}}(x;\kappa,\sigma,\theta)I(x > \theta)], \quad x \ge 0, \quad \mu,\kappa,\sigma,\theta > 0 \tag{8a}$$

and the CDF as

$$F_{\mathrm{HEG}}(x;\mathbf{P}_{\mathrm{HEG}}) = \frac{1}{Z}\{F_E(x;\mu)I(x \le \theta) + [F_E(\theta;\mu) + F_{\mathrm{GP}}(x;\kappa,\sigma,\theta)]I(x > \theta)\}. \tag{8b}$$

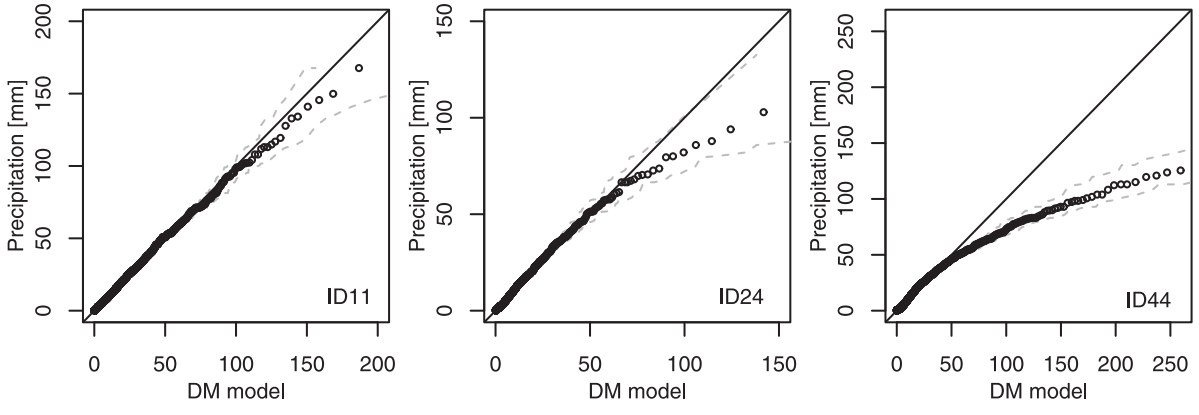

Figure 4. QQ plots of observed versus the DM distribution modeled precipitation quantiles of stations ID11 (left), ID24 (middle), and ID44 (right). The dash lines represent boundaries for the 95% confidence intervals.

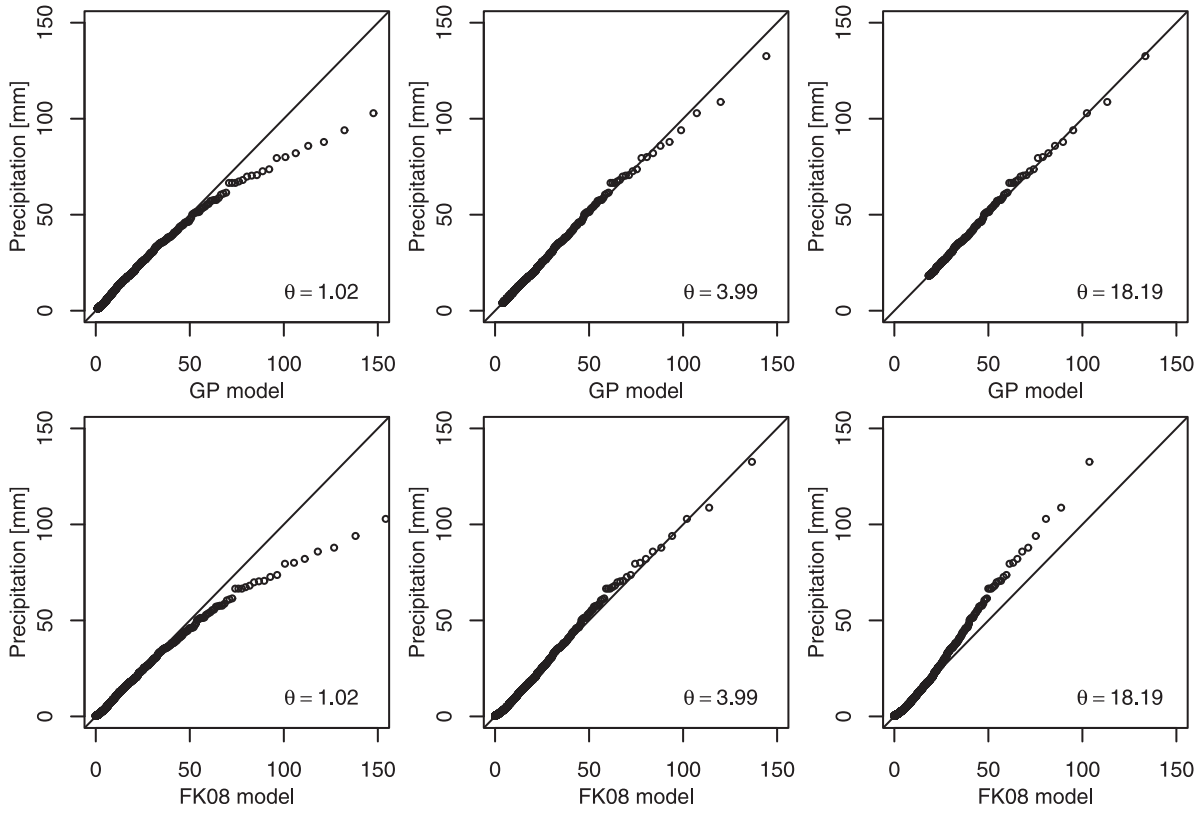

Figure 5. QQ plots of observed versus the GP distribution modeled quantiles of exceedances of precipitation over different thresholds (upper) and QQ plots of observed versus the FK08 distribution modeled full range of precipitation (lower) for station ID24.

The p-quantile function is

$$x_p = [-\mu \ln(1-pZ)]I(p \le F_E(\theta)) + \left[\theta + \frac{\sigma}{\kappa}\left(\left(pZ - 2 + \exp\left(-\frac{\theta}{\mu}\right)\right)^{-\kappa} - 1\right)\right] I(p > F_E(\theta)). \tag{8c}$$

The normalization constant Z assures that the hybrid density is integrated to one over its support and thus is given by

$$Z = \int_0^\theta f_E(x;\mu)dx + \int_\theta^\infty f_{\mathrm{GP}}(x;\kappa,\sigma,\theta)dx = F_E(\theta;\mu) + 1. \tag{8d}$$

CDFs of exponential and GP distributions are involved in the HEG distribution:

$$F_E(x;\mu) = 1 - \exp\left(-\frac{x}{\mu}\right), \tag{8e}$$

$$F_{\mathrm{GP}}(x;\kappa,\sigma,\theta) = 1 - \left(1 + \kappa\frac{x-\theta}{\sigma}\right)^{-\frac{1}{\kappa}}. \tag{8f}$$

To enforce the continuity of the hybrid density, i.e., $f_H(\theta+) = f_H(\theta-)$, the threshold θ is defined as the junction point of the exponential and the GP distributions and therefore can be explicitly expressed as a function of scale parameters of the two distributions, i.e.,

$$\theta = -\mu \ln\frac{\mu}{\sigma}. \tag{8g}$$

Apparently, the number of free parameter reduces to three. Thus, this distribution can be fully represented by $\mathbf{P}_{\mathrm{HEG}} = [\mu, \kappa, \sigma]$. Equation (8g) successfully bypasses the need for an explicit threshold selection. It is however cautioned that since only the PDF is forced to be continuous, it may converge to an unsmooth function, which nevertheless will not significantly influence the simulation and estimation. To remove such unsmoothness, one can force the derivative of the density to be continuous at the junction point. However, the flexibility of the distribution will decline.

[32] The HEG distribution adopts an exponential distribution for the low to moderate values, rather than a gamma distribution. This choice is made mainly for computational simplicity. It is recognized that directly learning the threshold θ by maximizing the log likelihood is challenging [*Frigessi et al.*, 2002; *Carreau and Bengio*, 2009]. A feasible approach is to learn it implicitly through expressing θ as a function of other parameters [*Carreau and Bengio*, 2009]. Taking the exponential distribution as a hybrid

component, threshold θ can be formulated as a closed function of the HEG parameters. However it is difficult to do so with the gamma distribution. In turn, it is infeasible to bypass the need for threshold selection in a supervised manner as by *Furrer and Katz* [2008]. The HEG distribution is aimed to offer an efficient and relatively simple option to model the full spectrum of precipitation amount. The price is the reduced flexibility in modeling the bulk of the data. This is a limitation of the proposed distribution. Fortunately, the divergence of performances between the exponential and gamma distributions is not too much for low to moderate values.

[33] One may note that in the DM distribution, data below the location parameter μ are modeled by the gamma distribution, and those above which are modeled by the GP distribution, when forcing τ to 0. It means that in this case the DM distribution reduces to an analog of HEG, i.e., hybrid gamma and GP distribution. The problem of this limiting distribution is the discontinuity at the location μ [*Furrer and Katz*, 2008]. A discontinuous density function is difficult to explain in practice, representing an unrealistic feature in precipitation [*Vrac and Naveau*, 2007]. Removing this discontinuity will again come across the difficulties as described above, assuming that one wants to avoid setting the threshold a priori.

[34] Another point worth noting is with respect to shape parameter κ, which determines the tail property of the HEG distribution. Negative, zero, and positive values of κ imply, respectively, bounded, light, and heavy tails. In the HEG distribution, κ is forced to be positive, considering the fact that daily precipitation is widely acknowledged being heavy tail distributed [*Koutsoyiannis*, 2004a, 2004b; *Vrac and Naveau*, 2007, *Furrer and Katz*, 2008]. The shape parameter κ can be restricted positive in one of two ways, with the use of a constrained optimization algorithm or by applying an exponential function to map the searching space of κ from real line onto positive real line. This study adopts the latter way. Moreover, we want to caution that in the case of zero κ, the GP distribution reduces to an exponential distribution and equation (8d) becomes incorrect.

5. Evaluation of Hybrid Distribution

[35] Since parameters of the hybrid distribution were estimated by the ML method and since the purpose is to model the full range of precipitation, one must answer the following two questions: (1) Is the ML estimator of the hybrid distribution asymptotically consistent and efficient? (2) Does the hybrid distribution perform better than or at least comparable to other distributions?

5.1. Asymptotic Property of the ML Estimator

[36] To answer the first question, random sample sets were generated with increasing size from the HEG distribution with parameters $\mathbf{P}_{\mathrm{HEG}} = [1.0, 0.2, 1.5]$, parameter estimates $\hat{\mathbf{P}}_{\mathrm{HEG}}$ were computed using the ML method for each sample set, and the asymptotic behavior of the ML estimator was empirically investigated. The sample size was increased with varying factor such that the asymptotic behavior was exhibited efficiently. For each sample size, random sampling and parameter estimation procedures were repeated 100 times to take into account the statistical variability.

5.1.1. Asymptotic Consistency of the ML Estimator

[37] The mean square error (MSE) between estimated and actual parameters was used to assess the asymptotically consistent property. If MSE decreased to zero as the sample size approached infinity, then the estimator was said to be asymptotically consistent. The MSE values of the ML estimates for each parameter are illustrated in the bar plot with increasing sample size, as shown in Figure 6, which indicates that MSE values for all parameters become smaller as the sample size increases. Observing the decreasing pattern of MSE, one can expect that as the sample size increases to a sufficiently large value MSE decays to zero or at least to a negligible value, which exactly meets the asymptotically consistent expectation of the ML estimators.

5.1.2. Asymptotic Efficiency of the ML Estimator

[38] The variance of estimates was used to indicate the asymptotically efficient behavior. If the variance of estimates decayed to zero as the sample size tended to infinity,

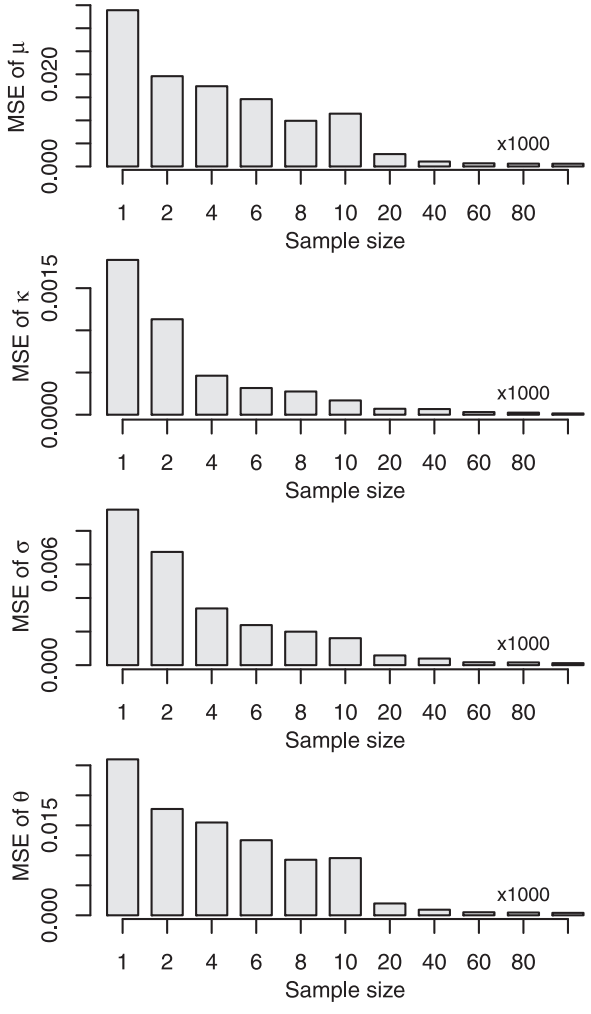

Figure 6. Behaviors of MSE of the ML estimators of the HEG distribution (parameterized by $\mathbf{P}_{\mathrm{HEG}} = [1.0, 0.2, 1.5]$) as the sample size increases.

then the estimator was considered asymptotically efficient. The distributions of parameters estimated by the ML method are shown by box plots in Figure 7. The span of the box plot shows a clear decreasing trend as the sample size increases. Note that when the sample size is large enough the estimated parameters are clustered within a very narrow interval centered or almost centered at real values. One can also expect that the interval will become much narrower and even negligible as the sample size goes to a sufficiently large value. This observation roughly indicates the asymptotic efficiency of the ML estimator since asymptotic efficiency involves that not only the variance of estimates decays to zero but also the rate at which it decays to zero.

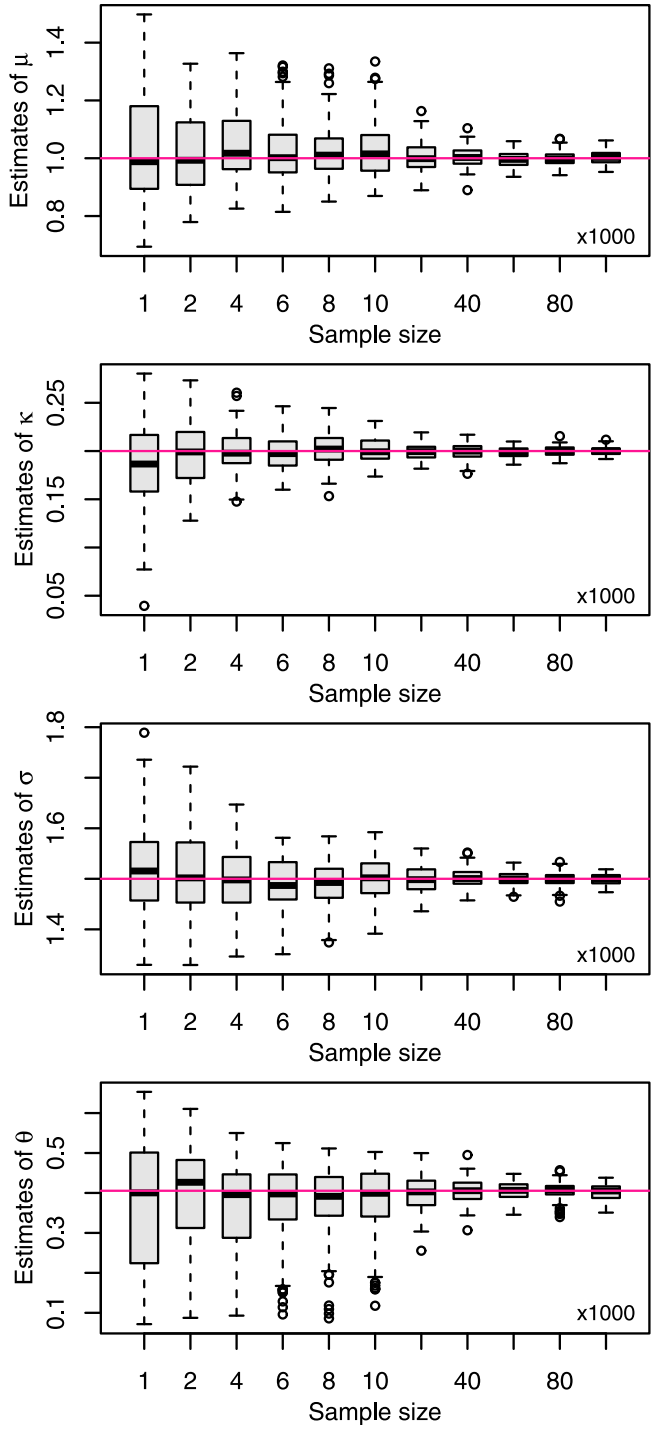

Figure 7. Distributions of the ML estimators of the HEG distribution (parameterized by $\mathbf{P}_{\mathrm{HEG}} = [1.0, 0.2, 1.5]$) as the sample size increases.

5.2. Preliminary Evaluation of Performance in Modeling Heavy Tail Data

[39] To answer the second question, the same scheme was used as in section 5.1. The parent distribution, however, was changed to an FK08 distribution parameterized as $\mathbf{P}_{\mathrm{FK08}} = [0.7, 17.4, 0.25, 3.9]$, which was obtained by fitting the model to the data from station ID24. The simulation experiment was conducted as follows: generate different training sets with increasing size and test set with a fixed size of 1000 from the parent FK08 distribution; fit each training set to the HEG distribution and other widely used candidates; compute the estimated probability densities for elements in the test set; and then compare them with the actual values, which can be calculated directly by the PDF of the FK08 distribution. Four candidate models were chosen: the HEG distribution, the ME distribution, the DM distribution, and the nonparametric KDE with Gaussian kernel.

[40] As in the previous simulation experiment, the size of training sets was increased with varying factor and for each size the above procedure was repeated 50 times. A relatively small number of repeats (50) was used mainly because of the high computation cost of the DM distribution. The FK08 distribution was excluded to avoid selecting the threshold in a supervised manner.

5.2.1. Evaluation of Performance Using Relative Log Likelihood

[41] To quantitatively assess different candidate models, the relative log likelihood (RLL) was computed as

$$R_l = -\frac{1}{1000}\sum_{i=1}^{1000} \log \frac{\hat{p}_i}{p_i}, \tag{9}$$

where p_i and $\hat{p}_i$, respectively, are the theoretical and estimated probability densities for elements in the test set. RLL is a measure of the divergence between the parent and the estimated distributions. The smaller the RLL is, the better the density estimator performs [*Carreau and Bengio*, 2009]. Summary statistics of RLL are listed in Tables 2, 3, 4, and 5, corresponding to the HEG distribution, the ME distribution, the DM distribution, and the KDE, respectively. The mean of RLL is a measure of the goodness-of-fit of the estimated density, whereas the coefficient of variation (CV) is a stability indicator.

[42] The number of failures for each model is also summarized in these tables. The DM distribution failed due to numerical instability caused by less accurate and sometimes false normalization constant. The resulting density was meaningless, for example, a complex value due to complex normalization constant, especially when the sample was small. The failure of KDE resulted when the estimated densities for large quantiles were too small compared with the true values in that the RLL tended to infinity.

Table 2. Summary Statistics of RLL for the HEG Distribution With Different Training Set Sizes *n*

n (×1000)	Failures	Min.	Mean	Max.	Cor. Coef.
1	0/50	0.00078	0.026	0.090	0.75
2	0/50	0.00025	0.023	0.155	0.93
4	0/50	0.0039	0.014	0.033	0.45
6	0/50	0.0023	0.014	0.030	0.44
8	0/50	0.0024	0.013	0.024	0.45
10	0/50	0.00052	0.013	0.030	0.47
20	0/50	−0.0027	0.011	0.024	0.54
40	0/50	0.00078	0.011	0.027	0.48
60	0/50	−0.00048	0.010	0.026	0.50
80	0/50	−0.0011	0.011	0.021	0.48
100	0/50	−0.00097	0.010	0.024	0.45

Table 4. Summary Statistics of RLL for the DM Distribution With Different Training Set Sizes *n*

n (×1000)	Failures	Min.	Mean	Max.	Cor. Coef.
1	20/50	0.0025	0.0217	0.062	0.70
2	13/50	0.0019	0.0126	0.033	0.58
4	3/50	−0.0023	0.0065	0.031	0.94
6	2/50	−0.0028	0.0039	0.0096	0.74
8	1/50	−0.0026	0.0050	0.019	0.87
10	0/50	−0.0026	0.0037	0.012	0.92
20	0/50	−0.0014	0.0014	0.0070	1.47
40	0/50	−0.0015	0.00088	0.0040	1.42
60	0/50	−0.0016	0.00054	0.0035	1.80
80	0/50	−0.0029	0.00040	0.0035	2.84
100	0/50	−0.00095	0.00056	0.0026	1.38

[43] Table 4 shows that the number of failures for the DM distribution was large, especially for small samples. When the sample size was less than 2000, the failure rate varied from 26% to 40%. The high failure rate indicates the existence of numerical problems due to its functional complexity. The failure of KDE especially appeared when the sample size was large since in this case more large values would emerge in the sample. This observation verified that the KDE was only reliable to model low to moderate values, but was, however, unable to appropriately simulate the upper tail behavior [*Markovich*, 2007; *Carreau and Bengio*, 2009].

[44] Based on the mean of RLL without considering the failure rate, the DM distribution ranked first, the HEG distribution second, the KDE third, and the ME distribution ranked fourth. Looking at the stability of these distributions as indicated by the CV values, the HEG distribution performed the best. The CV values for the HEG distribution were smaller than those of the DM distribution. Exceptions appeared when the sample size was less than 2000. This might be because of the high failure rate of the DM distribution. Only 30 and 37 of the 50 repeats were involved in the computation of CV, leading to suspicious and misleading results.

[45] From the view point of minimum RLL, the DM distribution seemed to be the best choice for heavy tailed data. It should be noted that in the DM model there are six free parameters, whereas only three free parameters are involved in the HEG distribution. Thus, it is argued that the small RLL is probably caused by over fitting.

5.2.2. Evaluation of Performance Using Information Criterion

[46] To further evaluate the distributions, the Akaike information criterion (*AIC*) [*Akaike*, 1974] and Bayesian information criterion (*BIC*) [*Schwarz*, 1978] were used. The information criteria take into account not only the goodness-of-fit but also the model complexity by penalizing the distribution with too many parameters. The smaller the *AIC* or *BIC*, the better the distribution.

[47] In this evaluation, the nonparametric KDE was excluded since it is broadly accepted that KDE is not suitable for heavy-tailed density estimation, which was also empirically verified by observations in section 5.2.1. The frequencies of selection of the three candidate models by *AIC* and *BIC* are summarized in Table 6. Results show that the DM distribution is penalized due to its functional complexity. And one can conservatively conclude that the small RLL values resulted from over fitting. For small samples, the ME distribution also performed well. This is not difficult to understand, because the majority of data in small samples were from the bulk of the distribution. As the sample size increased, more and more large quantiles were sampled and the inability of the ME distribution in modeling large values became apparent. Both *AIC* and *BIC* criteria indicated the HEG distribution to be the best choice.

5.3. Further Evaluation Using Precipitation Amount Records

[48] Does the preliminary simulation show that the HEG distribution is the best choice to model daily precipitation

Table 3 Summary Statistics of RLL for the ME Distribution With Different Training Set Sizes *n*

n (×1000)	Failures	Min.	Mean	Max.	Cor. Coef.
1	0/50	0.011	1.184	25.554	3.04
2	0/50	0.031	0.506	2.557	1.26
4	0/50	0.007	0.489	3.982	1.45
6	0/50	0.012	0.212	1.264	1.24
8	0/50	0.011	0.352	2.634	1.45
10	0/50	0.016	0.526	9.564	2.60
20	0/50	0.006	0.172	1.304	1.57
40	0/50	0.010	0.116	0.913	1.48
60	0/50	0.007	0.053	0.238	0.78
80	0/50	0.009	0.066	0.545	1.45
100	0/50	0.003	0.064	0.403	1.31

Table 5. Summary Statistics of RLL for the Gaussian KDE With Different Training Set Size *n*

n (×1000)	Failures	Min.	Mean	Max.	Cor. Coef.
1	0/50	0.027	0.126	0.361	0.58
2	0/50	0.017	0.0562	0.123	0.54
4	0/50	0.0023	0.0355	0.375	1.62
6	0/50	0.00029	0.0311	0.297	1.42
8	0/50	−0.0014	0.0223	0.0869	1.00
10	0/50	−0.0012	0.0371	0.628	0.09
20	0/50	−0.0012	0.0407	0.432	1.73
40	0/50	−0.00043	0.0767	0.749	2.03
60	0/50	−0.0011	0.0595	0.614	2.04
80	1/50	−0.0015	0.0444	0.526	2.19
100	2/50	−0.0009	0.0382	0.729	2.96

Table 6. Frequencies of Selections of the Three Parametric Candidate Models by *AIC* and *BIC* Obtained From Different Simulations

	HEG		ME		DM	
n (×1000)	*AIC*	*BIC*	*AIC*	*BIC*	*AIC*	*BIC*
1	22	22	28	28	0	0
2	31	31	19	19	0	0
4	41	41	9	9	0	0
6	45	46	5	4	0	0
8	47	47	3	3	0	0
10	46	46	4	4	0	0
20	49	49	1	1	0	0
40	50	50	0	0	0	0
60	50	50	0	0	0	0
80	50	50	0	0	0	0
100	50	50	0	0	0	0

amount? This question cannot be answered simply by only yes or no, at least at the current stage. That is because of the limited scope of the simulation study and the diverse distributional patterns of real precipitation, and also because the parameter estimation method influences the performance of the distribution. In section 5.3, different distributions were fitted to daily precipitation records from the 49 stations in order to examine: (1) if one can expect the HEG distribution to be widely used to model daily precipitation amount; and (2) are there any exceptions where this distribution does not fit the data well?

[49] The tail of the ME distribution was too light to capture extreme precipitation, as shown in the left panel of Figure 8. Although QQ plots for only three sample stations are shown here, this conclusion is valid for all other stations. Therefore, our concern would only focus on the two compound distributions. To quantitatively measure the goodness-of-fit, average distance (*AD*) from the scattered points to the 1:1 reference line in the QQ plot was exploited, i.e.,

$$AD = \frac{1}{N}\sum_{i=1}^{N} |x_i^{\mathrm{Obs}} - x_i^{\mathrm{Est}}| \sin\left(\frac{\pi}{4}\right). \tag{10}$$

Results showed that for only 11 of the 49 stations, *AD* values corresponding to the HEG distribution were greater than those of the DM distribution, even though the DM distribution has six free parameters rather than three for the

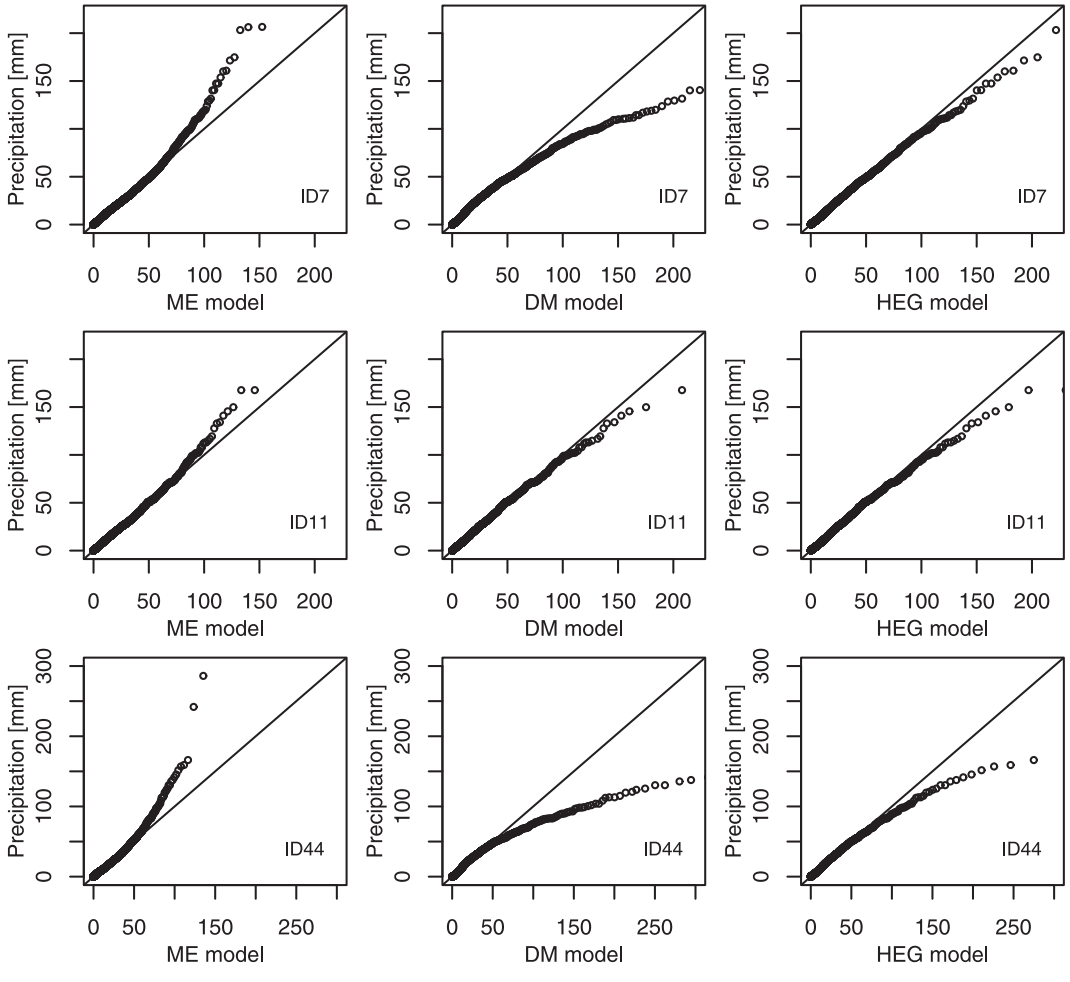

Figure 8. Representative patterns of the QQ plots of observed versus the ME distribution (left), the DM distribution (middle), and the HEG distribution (right) modeled quantiles of precipitation.

HEG distribution. As indicated from QQ plots, the larger *AD* values associated with the DM distribution are mainly caused by the over heavy tail.

[50] For further investigation, Figure 8 presents three representative patterns of QQ plots. The first pattern is shown in the top row panel, where the HEG distribution performs better than the DM distribution. In this case unsatisfactory goodness-of-fit of the DM distribution resulted mainly from the less accurate normalization constant. The second pattern is shown in the middle row panel, where both compound distributions fitted the data well. The DM distribution performed a little bit better over the HEG distribution with respect to minimum *AD* and maximum likelihood. This observation is expected, considering twice the number of parameters in the DM distribution. In this case, the *AIC* and *BIC* values were also computed. Neither of the two criteria provided information about over fitting, which agrees with the results of *Varc and Naveau* [2007]. It seems wise to choose the DM distribution if the problem of expensive computation is neglected. The presented HEG distribution is not intended, however, to replace the DM distribution but rather (1) to complement it in situations where it may have problems as at stations ID7 and ID44; and (2) to provide a relatively efficient, reliable, and simple way to model the full range of precipitation without losing any goodness-of-fit.

[51] The third pattern shown in the bottom panel of Figure 8 mainly signifies one disappointing fact that there are some situations in which both the DM distribution and the HEG distribution fail to capture extreme values. The noisy data, characterized by a few extremely large values appearing far away from the bulk of observations, is one of the reasons for this problem. Yet another probable reason is the misuse of a suboptimal parameter estimation method if one recalls that both upper tails of the two distributions are dominated by the GP distribution, which is outside of the exponential family. The ML estimator nicely performs only when the random variable belongs to the exponential family [*Castillo and Hadi*, 1995]. The following empirical study will provide justification for this possible explanation.

6. Other Parameter Estimation Approaches

[52] In the case that the ML estimator of the HEG distribution becomes invalid, how can one proceed? This is a common problem when working with the GP distribution, which is the tail component of the HEG distribution. Properly fitting the GP distribution has been approached by several researchers [*Hosking et al.*, 1985; *Castillo and Hadi*, 1995; *Singh and Guo*, 1995, 1997; *Castillo et al.*, 2005; *Luceno*, 2006; *Brazauska and Kleefeld*, 2009]. Two representative approaches are the elemental percentile (EP) method [*Castillo and Hadi*, 1995] and the maximum goodness-of-fit (MGF) method [*Luceno*, 2006]. The EP method is not an efficient option for the HEG distribution due to its somewhat cumbersome nature, which makes it difficult to express model parameters as functions of selected sample data and their empirical percentiles.

6.1. Two-Step Quantile Least Squares Method

[53] A two-step quantile least squares (QLS) method is described here. In the first step a number of samples are drawn with replacement from the original set and parameters are estimated based on the QLS method. For convenience, we denote these estimated parameter vectors as $\hat{\mathbf{P}}^1_{\mathrm{HEG}}, \hat{\mathbf{P}}^2_{\mathrm{HEG}}, \ldots, \hat{\mathbf{P}}^r_{\mathrm{HEG}}$, where r is the number of samples. The QLS estimates can be obtained as

$$\underset{\hat{\mathbf{P}}_{\mathrm{HEG}}}{\mathrm{Min.}} \quad \sum_{i=1}^{n} [x_{i:n} - \hat{x}_{i:n}]^2, \tag{11}$$

where $x_{i:n}$ and $\hat{x}_{i:n}$ are the observed and estimated ith order statistics, respectively. Since there is an explicit formula for the HEG quantile function, $\hat{x}_{i:n}$ can be computed from equation (8c), without resorting to any numerical method. The simple quantile function is another desirable property of the HEG distribution compared with the DM distribution in (1) straightforward random number simulation, and (2) multiple options for parameter estimation.

[54] Since the elemental QLS estimator is sensitive to samples, the second step is to obtain a final robust estimator by robust functions, say median function (MED). Thus the final estimator of $\mathbf{P}_{\mathrm{HEG}}$ can be expressed as

$$\hat{\mathbf{P}}_{\mathrm{HEG}}(MED) = \mathrm{median}(\hat{\mathbf{P}}^1_{\mathrm{HEG}}, \hat{\mathbf{P}}^2_{\mathrm{HEG}}, \hat{\mathbf{P}}^3_{\mathrm{HEG}}, \ldots, \hat{\mathbf{P}}^r_{\mathrm{HEG}}). \tag{12}$$

Goodness-of-fit analysis showed that QLS is a good alternative to the ML method when the latter has problems. However, the two-step QLS is a bootstrap based method. That means it is expensive in computation, which will deteriorate the computational efficiency advantage of the HEG distribution and thus be against one of our major goals, i.e., to search for an efficient way to simulate the full spectrum of precipitation.

6.2. Maximum Goodness-of-Fit Method

[55] The maximum goodness-of-fit (MGF) method is another alternative. The basic idea for the MGF method is to obtain parameter estimates by maximizing the goodness-of-fit of the distribution. The right-tail Anderson-Darling (RAD) statistic:

$$\begin{aligned} R_n^2 = {} & \frac{n}{2} - 2\sum_{i=1}^{n} F(\hat{\mathbf{P}}_{\mathrm{HEG}}; x_{i:n}) \\ & - \frac{1}{n}\sum_{i=1}^{n} (2i-1)\ln\left[1 - F(\hat{\mathbf{P}}_{\mathrm{HEG}}; x_{n=1-i:n})\right] \end{aligned} \tag{13}$$

is an efficient and robust choice for the HEG distribution. It is not surprising to assign more emphasis to the right tail of the distribution, considering the following two reasons. First, the bulk of the HEG is the exponential distribution which is easily fitted. Second, the tail is the GP distribution and the ML method may have problems to fit this part. In addition, QQ plots in Figure 8 also signify that the loss-of-fit always happened in the tail part.

[56] The MGF method was used to fit each precipitation set. Maximizing equation (13) was done with the use of the MATLAB function *fminsearch*. *AD* values were computed and compared with those obtained from the ML method, as given in Table 7. It is seen that *AD* reduced significantly after the MGF method was applied. The relative percentage

Table 7. Average Distance (*AD*) Values for the ML and MGF Fitted Distribution for Each Station

Station ID	*AD* (MLE)	*AD* (MGF)	Decrease Percentage (%)	Station ID	*AD* (MLE)	*AD* (MGF)	Decrease Percentage (%)
ID1	0.2566	0.2324	9.41	ID26	0.3075	0.2078	32.41
ID2	0.4370	0.2444	44.08	ID27	0.2373	0.1817	23.41
ID3	0.2553	0.1939	24.03	ID28	0.2028	0.1883	7.14
ID4	0.2480	0.2252	9.21	ID29	0.2808	0.2195	21.83
ID5	0.3231	0.2200	31.91	ID30	0.2987	0.2085	30.20
ID6	0.6048	0.3473	42.58	ID31	0.2949	0.2128	27.82
ID7	0.3015	0.1928	36.04	ID32	0.4531	0.2794	38.33
ID8	0.6097	0.3598	40.99	ID33	0.2079	0.2413	−0.1608
ID9	0.2605	0.2530	2.86	ID34	0.3681	0.2887	21.56
ID10	0.5075	0.3229	36.38	ID35	0.3001	0.2321	22.56
ID11	0.3296	0.2301	30.20	ID36	0.2644	0.2238	15.36
ID12	0.3127	0.2737	12.47	ID37	0.2853	0.1911	33.00
ID13	0.1850	0.1668	9.85	ID38	0.3182	0.2316	27.20
ID14	0.6366	0.4184	34.28	ID39	0.2181	0.1691	22.43
ID15	0.2711	0.1956	27.85	ID40	0.2924	0.2071	29.15
ID16	0.2558	0.1859	27.31	ID41	0.2897	0.2347	18.97
ID17	0.2872	0.2252	21.59	ID42	1.2676	0.2399	81.07
ID18	0.2804	0.2288	18.39	ID43	0.6539	0.4537	30.62
ID19	0.7450	0.4380	41.21	ID44	0.4850	0.2847	41.29
ID20	0.5139	0.1669	67.53	ID45	0.4140	0.3084	25.52
ID21	0.3761	0.2947	21.63	ID46	0.2521	0.2106	16.46
ID22	0.5521	0.3555	35.61	ID47	0.2323	0.1902	18.11
ID23	0.3701	0.2595	29.90	ID48	0.2326	0.1737	25.33
ID24	0.2733	0.1847	32.40	ID49	0.2120	0.1683	20.64
ID25	0.3133	0.2792	10.89				

of decrements ranged from 2.86% to 67.53%. Only one exception appeared at station ID33, where *AD* obtained by the ML method was smaller. Improvement of the goodness-of-fit can be more easily seen from Figure 9. The significant improvements empirically confirmed the aforementioned inference that the failure of the HEG distribution in capturing extremes was caused by the suboptimal parameter estimation method. Using the goodness-of-fit statistics (*AD*) for comparison is unfair for the ML method, because maximizing goodness-of-fit is the objective of the MGF method. The purpose here is to solve the problem of "where to proceed" when the ML method has problems. The MGF method does not intend to replace the ML method but rather to complement it in troublesome situations.

[57] Now one question arises: if the MGF method is sensitive to samples as it directly maximizes the agreement between the model and data. To answer this question, a simple simulation experiment was performed. A sample station, i.e., ID9, for which the ML estimator was optimal, was selected. The MFG method was used to fit nonzero observations. Then 500 random sets with fixed size, the same as the number of nonzero observations, were generated from the estimated HEG distribution. Finally, the MGF method was used to fit each set and its sensitivity was analyzed by assessing the variability of parameter estimates and quantile estimates. For comparison, the sample sets were also fitted using the ML method.

[58] Results are shown in Figure 10 by box plots. Compared with the ML estimator, the MGF estimator was more sensitive to the sample data, but not too much. Another point worth noting is that the mean of estimated parameters by the MGF is close to those of the true values. Therefore, it can be concluded that the MGF estimator is close to the ML estimator in the sense of root mean square error, even when the ML method is optimal. Real versus estimated quantiles were plotted in Figure 11. Similar patterns were found as those of the estimated parameters. The mean values of the estimated quantiles were almost the same as the theoretical values. The estimated quantiles by the MGF was also more sensitive to samples, but not as much as that in the parameter estimates. Put together, the above observations signify that the MGF estimator is a reliable alternative to the ML estimator, but with more variance.

6.3. Penalized Maximum Likelihood Method

[59] To take advantage of the less variance of the ML estimator and to reduce the likelihood of ML to converge to an over large shape parameter, the penalized maximum likelihood (PML) method is an appealing option. The PML method has been used for fitting extreme value distributions [*Coles and Dixon*, 1999]. The idea of the PML is to restrict the search space of the shape parameter by applying a penalty function. The adopted penalty function is of the form

$$p(\kappa) = \left(\exp\left(-\lambda\left(\frac{1}{(1-\kappa)} - 1\right)^{\alpha}\right)\right)^{\beta} \quad \kappa \in (0 \quad 1]. \tag{14}$$

As κ increases from 0 to 1, the penalty function decreases from 1 to 0. Compared with the one recommended by *Coles and Dixon* [1999], the additional parameter β provides more flexible control on the decreasing rate. After numerical experiments, it was found that the combination $\lambda = 0.1$, $\alpha = 0.8$, and $\beta = 0.15$ was suitable for the precipitation records reported in this study, as shown in Figure 12. The PML estimates are obtained by applying *fminsearch* to the penalized likelihood (or log likelihood), which is given by

$$L = \prod_{i=1}^{n} f_{\mathrm{HEG}}(x_i; \mathbf{P}_{\mathrm{HEG}}) p(\kappa). \tag{15}$$

[60] QQ plots of the PML fitted HEG distributions of sample stations were included in Figure 9. One can see that

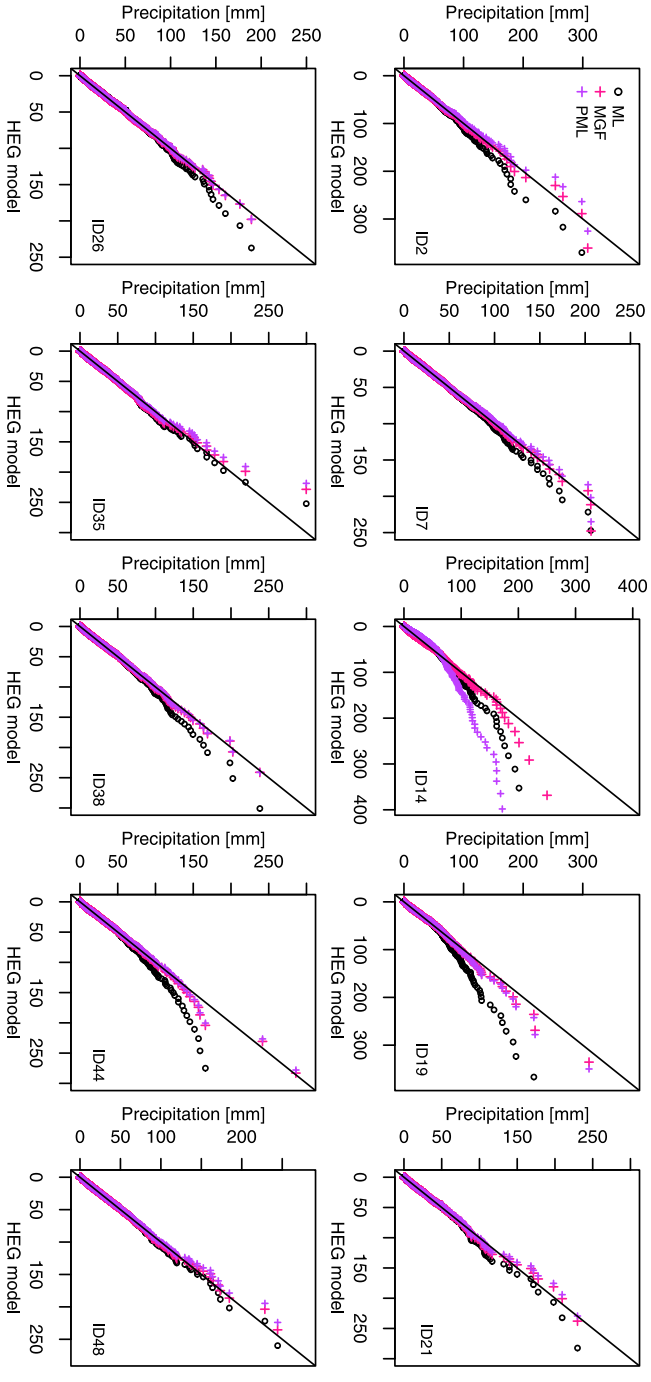

Figure 9. QQ plots of observed versus the HEG distribution modeled quantiles of precipitation using different model fitting methods.

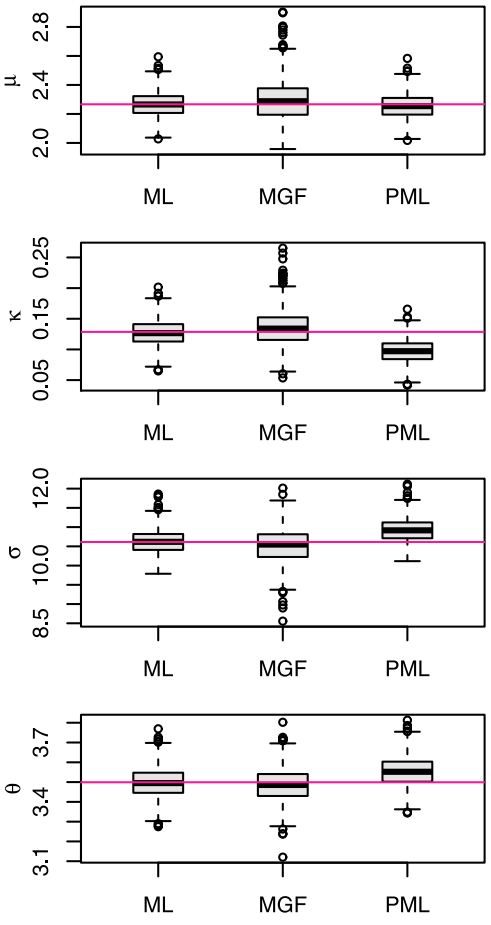

Figure 10. Box plots for HEG parameters fitted to 500 random samples by MLE, MGF, and PMLE methods, respectively. The random samples were drawn from the same parent distribution. True values are highlighted by deep pink lines.

PML can adequately fit the precipitation records. Only one exception among the 49 series appeared at station ID14, where the PML estimator performed worst among the three estimators. The same simulation experiment as for the sensitivity analysis of the MGF estimator was performed to investigate the influence of samples on the PML estimator. As expected, the PML estimator was less sensitive to samples than the MGF estimator and was comparable to the ML estimator, as presented in Figure 10. Similar observations were found in the quantile estimates, as shown in Figure 11. It is however cautioned that the reduced variance of the PML estimator is obtained at the expense of negatively biased shape parameter estimator, which was also remarked by *Coles and Dixon* [1999]. The negatively biased shape parameter will in turn lead to positively biased scale and location parameters, and negatively biased extreme quantiles. In terms of bias and variance, the PML estimator appears to be at least a competitive estimator to the other three and can be used as an alternative to the ML estimator.

[61] The described four estimating methods are all feasible options for fitting the HEG distribution, nevertheless, none of them being completely convincing in that it has been shown to be better than any other in every respect, for every data set. As to which method should be used, it involves a problem of tradeoff between bias and variance. In practice of precipitation simulation, we do not promote one estimator in favor of another, but recommend that both MGF and PML are reliable estimators in that generally they can provide adequate goodness-of-fit. Last but not least, we want to point out that the bootstrap based two-step estimating framework, as used in the two-step QLS

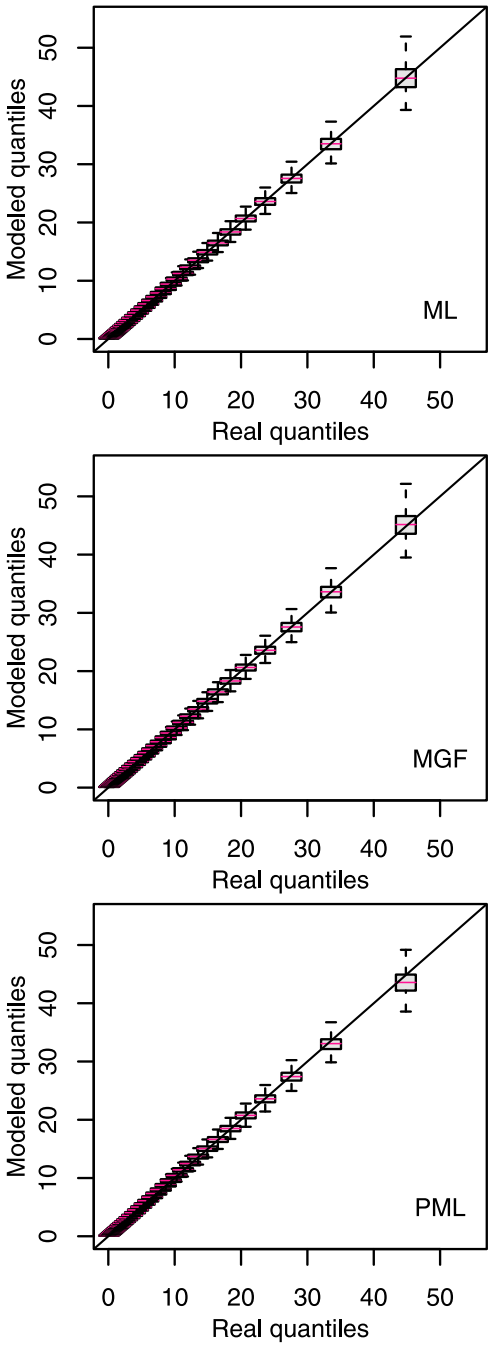

Figure 11. QQ plots of real and estimated quantiles by HEG distributions fitted to 500 random samples by MLE, MGF, and PMLE methods, respectively. The random samples were drawn from the same parent distribution. Boxes indicate the distribution of each estimated quantile.

estimator, might be used to reduce the variance of the MGF estimator, given that the computational efficiency is relatively less important. In this sense, the MGF method is preferred.

7. Conclusions

[62] This paper first examines the performance of existing distributions in modeling daily precipitation. Commonly

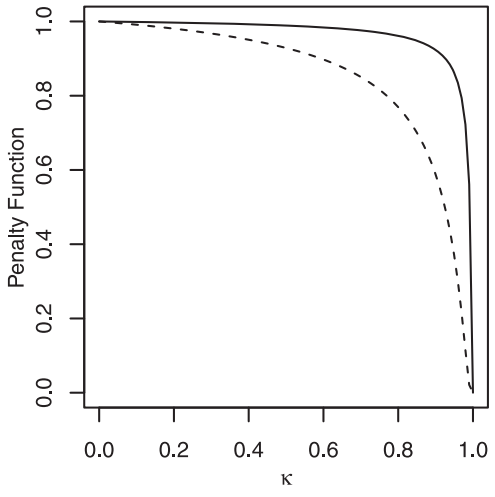

Figure 12. Penalty function of the shape parameter κ used in this study (solid line) and the one from *Coles and Dixon* [1999] (dashed line).

used single component distributions cannot realistically model extreme rainfall events in many cases but compound distributions can adequately do. However, the fitting of these compound distributions is plagued by several drawbacks, exemplified by functional complexity, numerical instability, data sensitivity, and supervised learning. In view of these drawbacks, we present a hybrid exponential and generalized Pareto distribution. This distribution is tested on 49 records across Texas. Results show that it is relatively simple and reliable in modeling the full spectrum of precipitation distribution. By expressing the threshold parameter of the generalized Pareto component in the hybrid model as a function of other parameters, the threshold can be implicitly learned in an unsupervised manner. Therefore the difficulty of threshold selection as in the conventional peak over threshold analysis can be circumvented. Moreover, attributing to its functional simplicity, random numbers of the hybrid distribution can be easily simulated by entering uniform random variates into its quantile function, which can be explicitly expressed.

[63] Parameters of the hybrid exponential and generalized Pareto distribution can be estimated using different methods, for instance maximum likelihood, two-step quantile least square, maximum goodness-of-fit, and penalized maximum likelihood methods. In most cases, the maximum likelihood estimator is an optimal choice because of its nice properties like statistical consistency and efficiency, and so on. The other three alternatives can be decent remedies when the maximum likelihood method is troublesome due to a few extremely large values appearing far away from the bulk of observations. As to which method should be used in practice, we do not promote one in favor of another, but recommend that both the maximum goodness-of-fit and the penalized maximum likelihood methods are reliable in that generally they can provide adequate goodness-of-fit for both the "bulk" and the "tail" of the data. The proposed hybrid exponential and generalized Pareto distribution might be incorporated into stochastic weather generators to provide an efficient way to realistically

simulate and downscale the full spectrum of daily precipitation. To ease the calculation effort and to reproduce results reported in this study, a suite of MATLAB functions is developed.

[64] There are several important issues appreciated in the rainfall modeling community that need more research in the future. One is to study the spatial distribution of the model parameters across a region so as to regionalize the model to be applicable at any location in the area. This issue is expected to be solved through a way inspired by *Wilks* [2008, 2009] and *Kleiber et al.* [2011]. On the other hand, all the reported analysis was limited within Texas. Evaluating the applicability and performance of the hybrid exponential and generalized Pareto distribution in diverse areas worldwide is another important task worthy of more efforts in the future.

Appendix

[65] Following *Frigessi et al.* [2002], the random numbers from the DM distribution can be sampled as follows:

[66] 1. Draw a uniform variate u from $U[0, 1]$.

[67] 2. If $u < 0.5$, then draw a random variate x from the Weibull distribution and evaluate the mixing function $p(x; \mu, \tau)$ at x, then retain x with probability $p(x; \mu, \tau)$ or reject it with probability $1 - p(x; \mu, \tau)$, and if so, resample again.

[68] 3. If $u \geq 0.5$, then draw a random variate x from the Pareto distribution and evaluate the mixing function $p(x; \mu, \tau)$ at x, then retain x with probability $p(x; \mu, \tau)$ or reject it with probability $1 - p(x; \mu, \tau)$, and if so, resample again.

[69] 4. Repeat step 1 to step 3 many times until a desired number of random variates have been sampled.

[70] The pseudo code for the sampling procedure following the rule of MATLAB is shown in Table A1.

Table A1. Pseudo Code for Random Numbers Simulation of the DM Distribution

```
Set a, b, μ, τ, κ, σ
while true
  u ← sample from uniform U[0,1]
  if 0.5 − u > eps
    x ← sample from gamma distribution parameterized by a and b
    p ← 1/2 + 1/π arctan((x−μ)/τ)
    r ← sample from uniform U[0,1]
    if 1 − p − r > eps
      R ← x
      return
    else
      continue
    end
  else
    x ← sample from Pareto distribution parameterized by κ and σ
    p ← 1/2 + 1/π arctan((x−μ)/τ)
    r ← sample from uniform U[0,1]
    if p − r ≥ eps
      R ← x
      return
    else
      continue
    end
  End
```

[71] **Acknowledgments.** This work was financially supported in part by the United States Geological Survey (USGS, Project ID: 2009TX334G) and TWRI through the project 'Hydrological Drought Characterization for TX under Climate Change, with Implications for Water Resources Planning and Management.' The constructive comments raised by the anonymous reviewers and the associate editor are gratefully acknowledged.

References

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Autom. Control AC-19*, *19*(6), 716–722.

Bárdossy, A., and E. Plate (1992), Space-time model for daily rainfall using atmospheric circulation patterns, *Water Resour. Res.*, *28*(5), 1247–1259.

Brazauska, V., and A. Kleefeld (2009), Robust and efficient fitting of the generalized Pareto distribution with actuarial applications in view, *Insurance: Math. Econ.*, *45*, 424–435.

Brissette, F. P., M. Khalili, and R. Leconte (2007), Efficient stochastic generation of multi-site synthetic precipitation data, *J. Hydrol.*, *345*, 121–133.

Carreau, J., and Y. Bengio (2009), A hybrid Pareto model for asymmetric fat-tailed date: The univariate case, *Extremes*, *12*, 53–76.

Castillo, E., and A. S. Hadi (1995), A method for estimating parameters and quantiles of distributions of continuous random variables, *Comput. Stat. Data Anal.*, *20*, 421–439.

Castillo, E., A. S. Hadi, N. Balakrishnan, and J. M. Sarabia (2005), *Extreme Value and Related Models with Applications in Engineering and Science*, John Wiley, Hoboken, NJ.

Chandler, R. E. (2005), On the use of generalized linear models for interpreting climate variability, *Environmetrics*, *16*, 699–715.

Coles, S. G. (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer, London.

Coles, S. G., and M. J. Dixon (1999), Likelihood-based inference for extreme value models, *Extremes*, *2*, 5–23.

Efron, B., and R. J. Tibshirani (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Frigessi, A., O. Haug, and H. Rue (2002), A dynamic mixture model for unsupervised tail estimation without threshold selection, *Extremes*, *5*, 219–235.

Furrer, E. M., and R. W. Katz (2007), Generalized linear modeling approach to stochastic weather generator, *Clim. Res.*, *34*, 129–144.

Furrer, E. M., and R. W. Katz (2008), Improving the simulation of extreme precipitation events by stochastic weather generators, *Water Resour. Res.*, *44*, W12439, doi:10.1029/2008WR007316.

Gabriel, K. R., and J. Neumann (1962), A Markov chain model for daily rainfall occurrence at Tel Aviv, *Q. J. R. Meteorol. Soc.*, *88*, 90–95.

Gomes, M., and O. Oliveira (2001), The bootstrap methodology in statistics of extremes-choice of the optimal sample fraction, *Extremes*, *4*(4), 331–358.

Hardwick Jones, R., S. Westra, and A. Sharma (2010), Observed relationships between extreme sub-daily precipitation, surface temperature, and relative humidity, *Geophys. Res. Lett.*, *37*, L22805, doi:10.1029/2010GL045081.

Harrold, T., A. Sharma, and S. J. Sheather (2003a), A nonparametric model for stochastic generation of daily rainfall occurrence, *Water Resour. Res.*, *39*(10), 1300, doi:10.1029/2003WR002182.

Harrold, T., A. Sharma, and S. J. Sheather (2003b), A nonparametric model for stochastic generation of daily rainfall amounts, *Water Resour. Res.*, *39*(12), 1343, doi:10.1029/2003WR002570.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood (1985), Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, *Technometrics*, *27*, 251–261.

Hundecha, Y., M. Pahlow, and A. Schumann (2009), Modeling of daily precipitation at multiple locations using a mixture of distributions to characterize the extremes, *Water Resour. Res.*, *45*, W12412, dio:10.1029/2008WR007453.

Hutchinson, E. F. (1995), Stochastic space-time weather models from ground based data, *Agric. For. Meteorol.*, *73*, 237–264.

Ison, N. T., A. M. Feyerherm, and L. Dean Bark (1971), Wet period precipitation and the Gamma distribution, *J. Appl. Meteor.*, *10*(4), 658–665.

Katz, R. W. (1974), Computing probabilities associated with the Markov chain model for precipitation, *J. Appl. Meteor.*, *13*(8), 953–954.

Katz, R. W. (1977), Precipitation as a chain-dependent process, *J. Appl. Meteor.*, *16*(7), 671–676.

Katz, R. W., and X. Zheng (1998), Mixture model for overdispersion of precipitation, *J. Climate*, *12*, 2528–2537.

Katz, R. W., M. B. Parlange, and P. Naveau (2002), Statistics of extremes in hydrology, *Adv. Water Resour.*, *25*, 1287–1304.

Kleiber, W., R. W. Katz, and B. Rajagopalan (2011), Daily spatio-temporal precipitation simulation suing latent and transformed Gaussian processes, *Water Resour. Res.*, *48*, W01523, doi:10.1029/2011WR011105.

Koutsoyiannis, D. (2004a), Statistics of extremes and estimation of extreme rainfall: 1. Theoretical investigation, *Hydrol. Sci. J.*, *49*(4), 575–590.

Koutsoyiannis, D. (2004b), Statistics of extremes and estimation of extreme rainfall: 2. Empirical investigation of long rainfall records, *Hydrol. Sci. J.*, *49*(4), 591–610.

Krstanovic, P. E., and V. P. Singh (1992), Evaluation of rainfall networks using entropy: I. theoretical development, *Water Resour. Manage.*, *6*, 279–293, doi:10.1007/BF00872281.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for time series resampling, *Water Resour. Res.*, *32*(3), 679–693.

Lenderink, G., and E. V. Meijgaard (2008), Increase in hourly precipitation extremes beyond expectations from temperature changes, *Nat. Geosci.*, *1*, 511–514, doi:10.1038/ngeo262.

Luceno, A. (2006), Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators, *Comput. Stat. Data Anal.*, *51*, 904–917.

Markovich, N. (2007), *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*, John Wiley, London.

Mehrotra, R., and A. Sharma (2007a), A semi-parametric model for stochastic generation of multi-site daily rainfall exhibiting low-frequency variability, *J. Hydrol.*, *335*, 180–193.

Mehrotra, R., and A. Sharma (2007b), Preserving low-frequency variability in generated daily rainfall sequences, *J. Hydrol.*, *345*, 102–120.

Mielke, P. W. (1973), Another family of distributions for describing and analyzing precipitation data, *J. Appl. Meteor.*, *12*(2), 275–280.

Mishra, A. K., and V. P. Singh (2010), Changes in extreme precipitation in Texas, *J. Geophy. Res.*, *115*, D14106, dio:10.1029/2009JD013398.

Naveau, P., M. Nogaj, C. Ammann, P. Yiou, D. Cooley, and V. Jomelli (2005), Statistical method for the analysis of climate extremes, *C. R. Geosci.*, *337*, 1013–1022.

Parlange, M. B., and R. W. Katz (2000), An extended version of the Richardson model for simulating daily weather variables, *J. Appl. Meteor.*, *39*(5), 610–622.

Rajagopalan, B., and U. Lall (1999), A k-nearest-neighbor simulator for daily precipitation and other variables, *Water Resour. Res.*, *35*(10), 3089–3101.

Richardson, C. W. (1981), Stochastic simulation of daily precipitation, temperature, and solar radiation, *Water Resour. Res.*, *17*(1), 182–190.

Roldan, J., and D. A. Woolhiser (1982), Stochastic daily precipitation models: 2. A comparison of distributions of amounts, *Water Resour. Res.*, *18*(5), 1461–1468.

Schoof, J. T., S. C. Pryor, and J. Surprenant (2010), Development of daily precipitation projections for the United States based on probabilistic downscaling, *J. Geophys. Res.*, *115*, D13106, doi:10.1029/2009JD013030.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464.

Sharma, A., and R. Mehrotra (2010), Rainfall generation, in *Rainfall: State of the Science*, edited by M. Gebremichael, pp. 215–246, American Geophysical Union, Washington, DC.

Singh, V. P. and H. Guo (1995), Parameter estimation for 2-parameter generalized Pareto distribution by the principle of maximum entropy, *Hydrol. Sci. J.*, *40*(2), 165–181.

Singh, V. P. and H. Guo (1997), Parameter estimation for 3-parameter generalized Pareto distribution by POME, *Water Resour. Manage.*, *9*, 81–93.

Solomon, S., D. Qin, M. Manning, M. Marquis, and others, Eds. (2007), *Climate Change 2007: The Physical Science Basis Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, New York.

Stern, D., and R. Coe (1984), A model fitting analysis of daily rainfall data, *J. R. Stat. Soc., Series A*, *147*, 1–34.

Todorovic, P., and D. A. Woolhiser (1975), A stochastic model for *n*-day precipitation, *J. Appl. Meteor.*, *14*(1), 17–24.

Vrac, M., and P. Naveau (2007), Stochastic downscaling of precipitation: From dry events to heavy rainfalls, *Water Resour. Res.*, *43*, W07402, doi:10.1029/2006WR005308.

Wilks, D. S. (1998), Multisite generalization of a daily stochastic precipitation generation model, *J. Hydrol.*, *210*, 178–191.

Wilks, D. S. (1999), Interannual variability and extreme-value characteristics of several stochastic daily precipitation models, *Agric. For. Meteorol.*, *93*, 153–169.

Wilks, D. S. (2008), High-resolution spatial interpolation of weather generator parameters using local weighted regressions, *Agric. For. Meteorol.*, *148*, 111–120, doi:10.1016/j.agrformet.2007.09.005.

Wilks, D. S. (2009), A gridded multisite weather generator and synchronization to observed weather data, *Water Resour. Res.*, *45*, W10419, doi:10.1029/2009WR007902.

Wilson, P. S., and R. Toumi (2005), A fundamental probability distribution for heavy rainfall, *Geophys. Res. Lett.*, *32*, L14812, doi:10.1029/2005GL022465.

Yan, Z., S. Bate, R. E. Chandler, V. Isham, and H. S. Wheater (2002), An analysis of daily maximum windspeed in northwestern Europe using generalized linear models, *J. Clim.*, *15*, 2073–2088.

Zheng, X., and R. W. Katz (2008a), Simulation of spatial dependence in daily rainfall suing multisite generators, *Water Resour. Res.*, *44*, W09403, doi:10.1029/2007WR006399.

Zheng, X., and R. W. Katz (2008b), Mixture model of generalized chain-dependent processes and its application to simulation of interannual variability of daily rainfall, *J. Hydrol.*, *349*, 191–199.

C. Li and A. K. Mishra, Department of Biological & Agricultural Engineering, Texas A&M University, College Station, TX 77843-2117, USA. (lichsunny@gmail.com)

V. P. Singh, Department of Biological & Agricultural Engineering and Department of Civil & Environmental Engineering, Texas A&M University, College Station, TX 77843-2117, USA.